Datasheets for Datasets

Motivation for Dataset Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

This dataset was created for the purpose of identifying successful betting strategies on the outcomes of NFL Regular Season and NFL Playoff games. Specifically, the goal of this dataset is to motivate people to build predictive models that can "beat the house."

What (other) tasks could the dataset be used for? Are there obvious tasks for which it should *not* be used?

This dataset should not be used to build predictive models that would be implemented outside of federal and state codes of law for sports betting.

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?

Previous models designed around this data are listed below (via Kaggle):
https://www.kaggle.com/twalters20/nfl-betting-model
https://www.kaggle.com/tobycrabtree/nflbetr
https://www.kaggle.com/northofmanhattan/james-hoffman-final-project-part-2
https://www.kaggle.com/northofmanhattan/james-hoffman-final-project-part-3

Who funded the creation of the dataset? If there is an associated grant, provide the grant number.

N/A

Any other comments?

Dataset Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

In nfl_stadiums.csv, the instances are the sports stadiums where an NFL game was played from 1966 to 2017. In nfl_teams.csv, the instances are the teams that were involved in an NFL game during the same time period. In spreadspoke_scores.csv, the instances are every single NFL game that has been played during the same time period.

Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)?

Each instance is its own independent entity, therefore there are no relationships between any instances.

How many instances of each type are there?

There are 100 instances in nfl_stadiums.csv, 41 in nfl_teams.csv, and 12.4k in spreadspoke_scores.csv.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances are related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

The following features are associated with the instances in nfl_stadiums.csv: stadium_name, stadium_location, stadium_open, stadium_close, stadium_type, stadium_address, stadium_weather_station_code, stadium_weather_type, stadium_capacity, stadium_surface, STATION, NAME, LATITUDE, LONGITUDE, ELEVATION.

The following features are associated with the instances in nfl_teams.csv: team_name, team_name_short, team_id, team_id_pfr, team_conference, team_division, team_conference_pre2002, team_division_pre2002.

The following features are associated with the instance in spreadspoke_scores.csv: schedule_date, schedule_season, schedule_week, schedule_playoff, team_home, score_away, score_away, team_away, team_favorite_id, spread_favorite, over_under_line, stadium, stadium_neutral, weather_temperature, weather_wind_mph, weather_humidity, weather_detail.

Is everything included or does the data rely on external resources? (e.g., websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version. Are there licenses, fees or rights associated with *any* of the data?

All data is included in the dataset.

Are there recommended data splits or evaluation measures? (e.g., training, development, testing; accuracy/AUC)

N/A.

What experiments were initially run on this dataset?

Have a summary of those results and, if available, provide the link to a paper with more information here.

The previous models built using this dataset are listed below with their accompanying results (if available):

Model1: https://www.kaggle.com/twalters20/nfl-betting-model
Results: https://tywalters97.wixsite.com/home/single-post/2018/10/20/NFL-Betting-on-Game-Results
Model 2: https://www.kaggle.com/tobycrabtree/nflbetr
Model 3:https://www.kaggle.com/northofmanhattan/james-hoffman-final-project-part-2
Model 4: https://www.kaggle.com/northofmanhattan/james-hoffman-final-project-part-3

Any other comments?

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software pro- gram, software interface/API; how were these con- structs/measures/methods validated?)

The data was collected off of various websites using a scraping tool listed on the Kaggle page as spreadspoke.R. NFL data came from public websites such as ESPN, NFL.com, and Pro Football Reference. Weather data is from NOAA data. Betting data is from http://www.repole.com/sun4cast/data.html for 1978-2013 seasons. From 2013 on betting data reflects lines available at sportsline.com.

Who was involved in the data collection process? (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)

N/A.

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame?

N/A.

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part of speech tags; model-based guesses for age or language)? If the latter two, were they validated/verified and if so how?

All of the data in these datasets were directly observable information.

Does the dataset contain all possible instances? Or is it, for instance, a sample (not necessarily random) from a larger set of instances?
These datasets contain all possible instances of NFL games played between 1966 and 2017.

If the dataset is a sample, then what is the population?
What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geographic coverage)? If not, why not (e.g., to cover a more diverse range of in- stances)? How does this affect possible uses?
N/A.

Is there information missing from the dataset and why?
(this does not include intentionally dropped instances; it might include, e.g., redacted text, withheld documents) Is this data missing because it was unavailable?
There is missing information from all three datasets. This missing information is mainly in the features associated with weather data, which were likely unavailable data for games within specific time frames.

Are there any known errors, sources of noise, or redundancies in the data?
N/A.

Any other comments?

Data Preprocessing

What preprocessing/cleaning was done? (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values, etc.)
The datasets were built using data pulled from various websites listed on the main Kaggle page.

Was the "raw" data saved in addition to the preprocessed/cleaned data? (e.g., to support unanticipated future uses)
No.

Is the preprocessing software available?
No.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?
N/A.

Any other comments?

Dataset Distribution

How is the dataset distributed? (e.g., website, API, etc.; does the data have a DOI; is it archived redundantly?)

The data is distributed through Kaggle's Datasets collections. The link for this data specifically is located at https://www.kaggle.com/tobycrabtree/nfl-scores-and-betting-data.

When will the dataset be released/first distributed? (Is there a canonical paper/reference for this dataset?)

N/A.

What license (if any) is it distributed under? Are there any copyrights on the data?

N/A.

Are there any fees or access/export restrictions? Any other comments?

N/A.

Any other comments?

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?

How does one contact the owner/curator/manager of the dataset (e.g. email address, or other contact info)?

The data is owned and is being maintained by the Kaggle user *spreadspoke*. There is no specific contact information listed for this user, but the user is very active in the Discussion thread on the datasets page location.

Will the dataset be updated? How often and by whom? How will updates/revisions be documented and communicated (e.g., mailing list, GitHub)? Is there an erratum?

There is no information regarding future updates.

If the dataset becomes obsolete how will this be communicated?

This information will likely come through the same Kaggle page it is currently located on.

Is there a repository to link to any/all papers/systems that use this dataset?

https://www.kaggle.com/tobycrabtree/nfl-scores-and-betting-data/kernels

If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?
Suggestion can be made through Kaggle's Discussion thread.

Any other comments?

Legal and Ethical Considerations

If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection? (e.g., datasets that collect writing, photos, interactions, transactions, etc.)

N/A.

If it relates to other ethically protected subjects, have appropriate obligations been met? (e.g., medical data might include information collected from animals)

N/A.

If it relates to people, were there any ethical review applications/reviews/approvals? (e.g. Institutional Review Board applications)

N/A.

If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications? If consent was obtained, how? Were the people provided with any mechanism to revoke their consent in the future or for certain uses?

N/A.

If it relates to people, could this dataset expose people to harm or legal action? (e.g., financial social or otherwise) What was done to mitigate or reduce the potential for harm?

N/A.

If it relates to people, does it unfairly advantage or dis- advantage a particular social group? In what ways? How was this mitigated?

N/A.

If it relates to people, were they provided with privacy guarantees? If so, what guarantees and how are these ensured?

N/A.

Does the dataset comply with the EU General Data Protection Regulation (GDPR)? Does it comply with any other standards, such as the US Equal Employment Opportunity Act?

N/A.

Does the dataset contain information that might be considered sensitive or confidential? (e.g., personally identifying information)

N/A.

Does the dataset contain information that might be considered inappropriate or offensive?

N/A.

Any other comments?